Research Article
Open Access

# Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis

## Abstract

Many applied disciplines report treatment means and their standard errors (or confidence intervals) in the presentation of experimental results. Often, overlapping standard error bars (or confidence intervals) are used to draw inferences about statistical differences (or lack thereof) between treatments. This practice can lead to misinterpretations about treatment effects because it lacks type I error rate control associated with formal hypothesis tests. Theoretical considerations and Monte Carlo methods are used to show that the probability that standard error bars overlap is affected by heterogeneity of variances in an unpaired data collection setting; in a paired setting, this probability is affected by heterogeneity of variances, degree and direction of non—independence (covariance) between treatments, and the variance of random pairing effects. As a result, basing inferences on overlapping standard error bars is a decision tool with type I error rates ranging from 16% to 32% in an unpaired setting, and from 0% to 32% in a paired setting. In contrast, type I error rates associated with overlapping 95% confidence intervals are at most 5% and generally much smaller. These considerations apply to one— and two—sided tests of hypotheses. In multivariate applications, non—overlapping hypothesis and error ellipses are reliable indicators of treatment differences.

**Keywords:** *P* value, treatment mean comparison, type I error rate

Volume 7 Issue 1 - 2018

David B Wester
Caesar Kleberg Wildlife Research Institute, Texas A&M University—Kingsville Kingsville, USA

**Correspondence:** David B Wester, Caesar Kleberg Wildlife Research Institute, Texas A&M University—Kingsville Kingsville, Texas, USA, Email david.wester@tamuk.edu

**Received:** January 21, 2018 | **Published:** February 26, 2018

## Introduction

There are several different philosophical/methodological approaches commonly used in the applied disciplines to investigate effects under study in designed experiments. In a letter to Jerzy Neyman (12 February 1932), Fisher[1] wrote, ". . . the whole question of tests of significance seems to me to be of immense philosophical importance." And in 1929, Fisher[2] wrote:

> "In the investigation of living beings by biological methods statistical tests of significance are essential. Their function is to prevent us being deceived by accidental occurrences, due not to causes we wish to study, or are trying to detect, but to a combination of many other circumstances which we cannot control."

Little[3] provided a conceptual history of significance tests and Stephens et al.[4] documented its prevalence in ecology and evolution. Salsburg[5] wrote that, '. . . hypothesis testing has become the most widely used statistical tool in scientific research,' and Lehmann[6] suggested that *t* and *F* tests of hypotheses 'today still constitute the bread and butter of much of statistical practice.' Criticism of null hypothesis significance testing also has a long history[7]. Robinson & Wainer[8] observed that as the number of criticisms of this approach has increased so has the number of defenses of its use increased.[4,9]

In addition to null hypothesis testing, the importance of 'effect size' was recognized early as well. Deming[10] wrote that, 'The problem is not whether differences [between treatments] exist but how great are the differences . . . .' Numerous authors have encouraged more attention to effect size estimation.[11] Bayesian methods[12] that include credibility intervals[13] and multi—model selection[14] also have been offered as alternatives to null hypothesis testing.

Each of these approaches has strengths and weaknesses which should be appreciated in order to most effectively use them to address a research hypothesis—and it is not the intent of this paper to enter into this discussion. My goals are more modest: to clarify that

i. Use of overlapping standard error bars (or confidence intervals) estimated around each of two means is not the same thing as estimating a confidence interval around a difference between two means, and to suggest that

ii. If an assessment of statistical significance is desired, then a formal test of hypothesis should be used rather than an assessment of overlap of standard errors bars and/or confidence intervals. Although statisticians are well aware of these issues, I show (below) that many applied scientists in a variety of disciplines are less clear.

Consider an experiment designed to compare weight change of animals on several different rations. If animals are relatively homogeneous (similar initial weight, etc.) then a completely randomized design can be used; alternatively, variability in initial weight can be blocked out in a randomized block design. In either case, a significant *F* test on treatment mean equality often is followed up with pairwise comparisons among treatments.

One of the most common ways to present results from an experiment designed to compare two or more treatments is to graph or tabulate their means. Presenting treatment means without also presenting some measure of the variability associated with these estimates, however, is generally considered 'telling only half of the story.' It is usually recommended, therefore, to present treatment means together with their standard errors or with confidence intervals. Interpretation of these measures of variability must be done carefully in order to properly understand the effect of the treatments.

Data from these kinds of studies can be analyzed and presented in several ways (Figure 1). We might report that mean weight change in the control diet was $\overline{Y}_C$ = 11 lbs whereas mean weight change in the experimental diet was $\overline{Y}_E$ = 25 lbs (Figure 1a). If only treatment means (i.e., *point* estimates) are presented e.g.,[15], critical information—in particular, a measure of the variability associated with these estimates—is missing. Clearly, the findings of this experiment can be enriched with a more complete presentation.

Consider what is gained by providing standard errors together with treatment means (Figure 1b). If we reported $\overline{Y}_C$ = 11 ± 1.1 lbs and $\overline{Y}_E$ = 25 ± 3.6 lbs, our colleagues likely would nod in approval. In contrast, if we reported $\overline{Y}_C$ = 11±9 and $\overline{Y}_E$ = 25±13 lbs, the high variability might call into question our experimental technique, or at least raise questions about the homogeneity of our experimental material—and providing standard errors allows a reader to make such assessments. So far, this approach involves providing standard errors for descriptive purposes only.

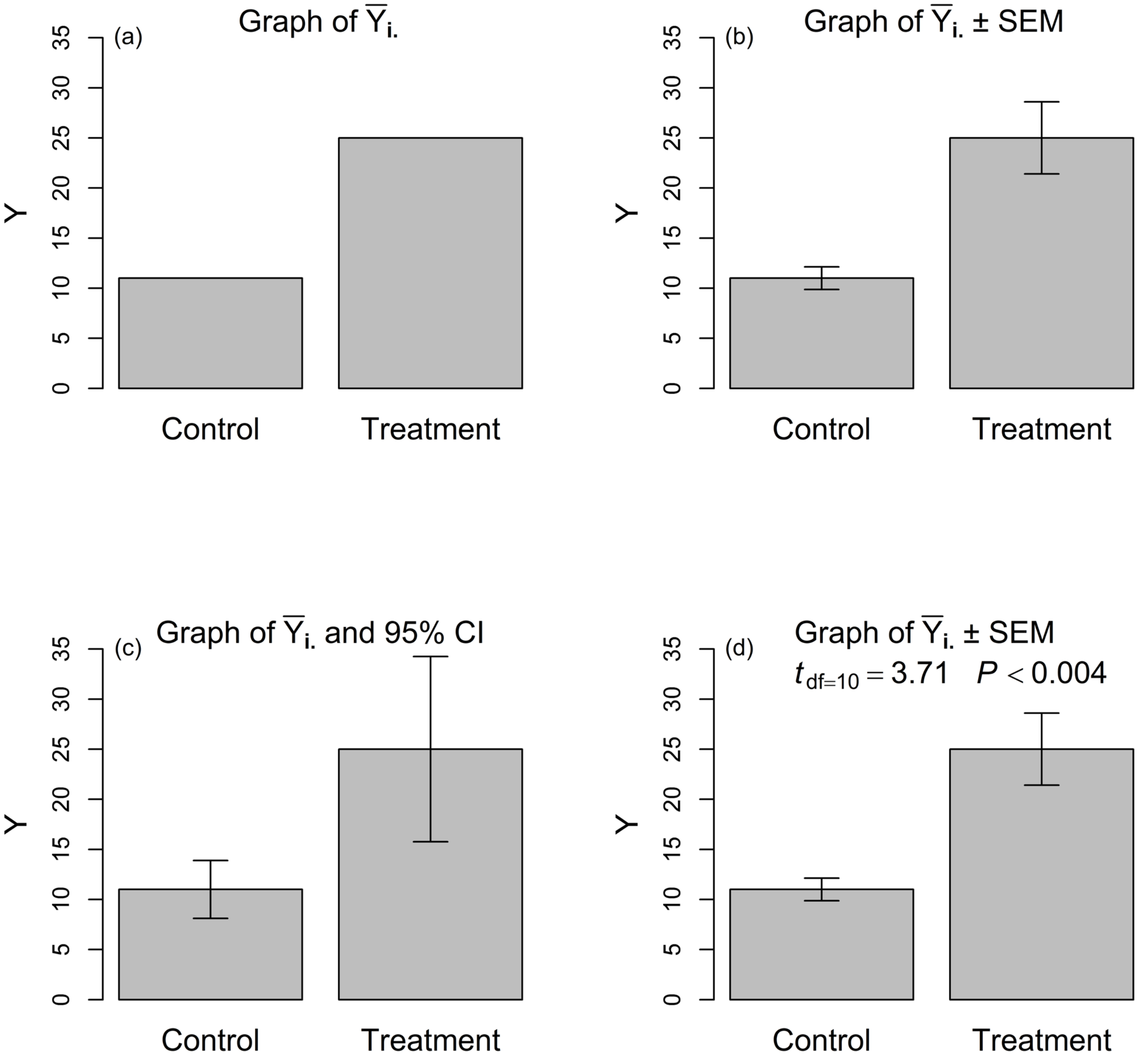

**Figure 1** Four ways to present results from an experiment designed to compare the equality of means in two treatments: (a) treatment means only; (b) treatment means and standard error bar for each mean; (c) treatment means and confidence interval for each mean; (d) treatment means, standard error bar for each mean, test statistic (with *df*) for hypothesis of treatment mean equality, and *P*-value associated with test statistic.

However, some researchers draw conclusions about treatment differences based on whether or not the standard errors around the treatment means overlap. For example, Beck et al.,[16] wrote, 'Non—overlapping standard errors were used as evidence of significant differences . . . among [treatments].' Jones et al.,[17] wrote, 'Significant difference . . . between [treatments] . . . was determined by non—overlapping standard errors.' Wisdom & Bate[18] wrote that '. . . we found no difference (overlapping standard errors) . . . [between treatments].' And Dwan´isa et al.[19] wrote that 'It bears mentioning that the error bars overlap thus indicating that [the response variable] may be the same in both [treatments].' Huang et al.[20] wrote that 'we remind the reader that for two [treatments] whose standard errors overlap, one may conclude that their difference is not statistically significant at the 0.05 level. However, one *may not conclude*, in general, that [treatments] whose standard errors do not overlap are statistically different at the 0.05 level.' Finally, some authors evidently consider a non—significant analysis of variance *F* test and overlapping standard error bars as synonymous: 'Each of the ANOVA models. . . was non—significant, bearing out the impression of stability suggested by overlapping standard error bars around these means'.[21]

A little reflection suggests an immediate difficulty: if our standard errors do not overlap and we choose to declare a difference between treatments, what '*P*' value should we attach to our inference? After all, there is nothing in the calculation of two treatment means and their standard errors that involves any use of the usual tabular *t* or *F* distributions.

Some researchers present treatment means and corresponding confidence intervals estimated for each mean. In our example, we might report that $\overline{Y}_C$ = 11 (95% CI: 8.1, 13.9) and $\overline{Y}_E$ = 25 (95% CI: 15.7, 34.3) (Figure 1c). Of course, a confidence interval is simply a treatment mean with a "± 'scaled' standard error" (where the scaling factor is the appropriate value from Student's *t* distribution), but it carries with it the added inferential information of an *interval* estimate—which can be adapted to the researcher's needs by the selection of an alpha level—about each treatment mean that the standard error alone lacks.

As with standard errors, some scientists compare treatments by examining the overlap of confidence intervals estimated around each treatment. This inference may be only implicit. For example, 'Estimated species richness was 38 (95% CI: 30.1—46.2) at undiked sites compared to 33 (95% CI: 29.2—36.7) at diked wetlands, but the confidence intervals overlapped considerably';[22] 'Despite the reduction in coyote predation rate during 2010, non—coyote—predation rate did not increase (95% CIs overlapped extensively) . . .'.[23] More explicitly, Mantha et al.[24] wrote, 'When two means based on independent samples with more than 10 observations in each group are compared, and the 95% CIs for their means do not overlap, then the means are significantly different at *P* < 0.05.' Bekele et al.,[25] applied this approach to medians: 'When the 95% confidence intervals of the medians overlap, the medians being compared were considered statistically not different at 5% probability level; otherwise the medians were considered statistically different.' In contrast to the approach of comparing the overlap of standard errors, we will see that this approach is actually quite conservative, with a type I error rate generally much lower than 5% when 95% confidence intervals are estimated.

The most rigorous and complete analysis and presentation provides treatment means, their standard errors (or confidence intervals), and results from a formal test of hypothesis of treatment mean equality (Figure 1d); we would report that '$\overline{Y}_C$ = 11 ± 1.1 lbs and $\overline{Y}_E$ = 25 ± 3.6 lbs, $t_{df=10} = 3.72$, $P < 0.004$.' This is the only inferential approach (for our experimental designs) that operates at the nominal alpha level. This approach is equivalent to estimating a confidence interval around the difference between two treatment means and comparing it to a value of zero.

In this paper, I explore the interpretation of overlapping standard errors, overlapping confidence intervals, and tests of hypothesis in univariate settings to compare treatment means. These principles, usually laid down in introductory courses in biometry, biostatistics and experimental design, are of course understood by statisticians. As the selected quotations above illustrate, however, these basic principles are often overlooked or misunderstood by applied scientists, and the resulting confusion about what is 'significant' (or not) in an experiment can have far—reaching implications in the interpretation of scientific findings. Simulation methods are used to illustrate key concepts. Results of Payton et al.[26,27] for unpaired settings are confirmed; new results are provided for paired samples which also incorporate the influence of random effects (resulting from pairing); one—sided hypothesis tests are considered; and experiment—wise error rates for experiments with more than two treatments are considered.

## Methods

For each of *N* = 10, 000 simulated data sets, pseudo—random samples of size *n* = 5, 10, 20, 30, 50, 100, or 1,000 were generated for each of *t* = 2 populations (treatments) for which $\mu_1 = \mu_2$. For a completely randomized design, the linear (fixed effects) model is

$$Y_{ij} = \mu + \tau_i + e_{(i)j} \qquad (1)$$

$i = 1, 2; j = 1, 2, \ldots, n$; and $\tau_i = 0$ when $\mu_1 = \mu_2$.

Population standard deviations were set at (1) $\sigma_{e_1} = \sigma_{e_2} = 4$, or (2) $\sigma_{e_1} = 4, \sigma_{e_2} = 16$. A $(1-\alpha)100\%$ confidence interval around a treatment mean is $\overline{Y}_{i.} \pm t_{\alpha/2,(n-1)} \frac{\hat{\sigma}_{e_i}}{\sqrt{n}}$, where $\overline{Y}_{i.} = \sum_j^n Y_{ij} / n$ and $\hat{\sigma}_{e_i} = \sqrt{\sum_j^n \frac{\left(Y_{ij} - \overline{Y}_{i.}\right)^2}{n-1}}$. For a randomized block design, the linear (mixed effects) model is

$$Y_{ij} = \mu + b_j + \tau_i + e_{(ij)} \qquad (2)$$

$i = 1, 2; j = 1, 2, \ldots, n$; and $\tau_i = 0$ when $\mu_1 = \mu_2$; the random effect, $b_j \sim N\left(0, \sigma_b^2\right)$, is responsible for pair—to—pair variation (i.e., the random 'block' effect). Population standard deviations were set at $\sigma_{e_1} = \sigma_{e_2} = 4$; population covariances between treatments were set at $\sigma_{e_{12}} = 0, 4 \text{ or} -1$; and $\sigma_b^2 = 0, 4 \text{or } 16$. A $(1-\alpha)100\%$ confidence interval around a treatment mean is $\overline{Y}_{i.} \pm t_{\alpha/2,(n-1)} \sqrt{\left(\hat{\sigma}_{e_i}^2 + \hat{\sigma}_b^2\right)/n}$ ,

**Citation:** Wester DB. Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis. *Biom Biostat Int J.* 2018;7(1):73–85. DOI: 10.15406/bbij.2018.07.00192

where $\bar{Y}_{i.} = \sum_j^n Y_{ij} / n$ and $\hat{\sigma}_{e_i}^2 + \hat{\sigma}_b^2 = \sum_j^n \frac{\left(Y_{ij} - \bar{Y}_{i.}\right)^2}{n-1}$.

The proportions of standard errors bars and confidence intervals that overlapped were counted in the $N$ = 10, 000 simulated data sets for each experimental design. Also, for each data set, the hypothesis $H_0 : \mu_1 = \mu_2$ *vs* $H_0 : \mu_1 \neq \mu_2$ was tested. For the completely randomized design, this hypothesis was tested with an unpaired *t* test: $t_c = \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{\sqrt{2\,MSE_{CRD} / n}}$, where $MSE_{CRD} = \sum_i^2 \sum_j^n \frac{\left(\bar{Y}_{ij} - \bar{Y}_{i.}\right)^2}{t\left(n-1\right)}$ is the error mean square from an analysis of variance with expectation $E[MSE_{CRD}] = \sigma_e^2 = \left(\sigma_{e_1}^2 + \sigma_{e_2}^2\right)/2$, the pooled error variance[28]; $\left|t_c\right|$ was compared to a tabular Student's *t* value, $t_{\alpha/2;\, t(n-1)}$ for the case of homogeneous variances and to $t_{\alpha/2;v'}$ where $v' = \frac{\left(\hat{\sigma}_{e_1}^2 / n + \hat{\sigma}_{e_2}^2 / n\right)^2}{\left(\hat{\sigma}_{e_1}^2 / n\right)^2 / \left(n-1\right) + \left(\hat{\sigma}_{e_2}^2 / n\right)^2 / \left(n-1\right)}$ [29] for the case of heterogeneous variances. For a randomized block design, this hypothesis was tested with $t_c = \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{\sqrt{2\,MSE_{RBD} / n}}$, where $MSE_{RBD} = \sum_i^2 \sum_j^n \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{j.} + \bar{Y}_{..}\right)^2 / \left[\left(t-1\right)\left(n-1\right)\right]$ is the error mean square for the analysis of variance with expectation $E\left[MSE_{RBD}\right] = \bar{\sigma}_{e.}^2 - \sigma_{e_{12}}$ [30] where $\bar{\sigma}_{e.}^2 = \left(\sigma_{e_1}^2 + \sigma_{e_2}^2\right)/2$ ; $\left|t_c\right|$ was compared to a tabular *t* value, $t_{\alpha/2;(t-1)(n-1)}$. (The difference in notation—$\sigma_e^2$ for a CRD and $\bar{\sigma}_{e.}^2$ for an RBD—reflects the different assumptions for these two experimental designs: in a CRD variances within treatments are assumed to be homogenous and are pooled whereas in an RBD variances within treatments are not assumed to be homogenous and are averaged.)

### Treatment means were compared using three distinct approaches:

**Approach A:** standard errors were estimated for each treatment mean and inferences were based on the proportion of overlapping standard errors from the $N$ = 10, 000 simulated data sets.

**Approach B:** (1−α)100% confidence intervals were estimated for each treatment mean and inferences were based on the proportion of overlapping confidence intervals from the $N$ = 10, 000 simulated data sets and

**Approach C:** a test of the hypothesis, hypothesis $H_0 : \mu_1 = \mu_2$ *vs* $H_0 : \mu_1 \neq \mu_2$. For the CRD, an unpaired *t* test was used with $t\left(n-1\right)$ error *df* when variances were homogeneous and Satterthwaite's adjustment when variances were heterogeneous; for the RBD, a paired *t* test with $\left(t-1\right)\left(n-1\right)$ error *df* was used.

## Results

### Theoretical considerations

For an unpaired setting (eq. 1), the probability that confidence intervals around two treatment means overlap is Payton et al.[27]:

$$P\left(\text{overlap}\right) = 1 - P\left[\bar{Y}_{1.} + t_{\alpha/2;(n-1)}\left(\hat{\sigma}_{e_1} / \sqrt{n}\right) < \bar{Y}_{2.} - t_{\alpha/2;(n-1)}\left(\hat{\sigma}_{e_2} / \sqrt{n}\right)\right] -$$

$$P\left[\bar{Y}_{2.} + t_{\alpha/2;(n\text{-}1)}\left(\hat{\sigma}_{e_2} / \sqrt{n}\right) < \bar{Y}_{1.} \text{-} t_{\alpha/2;(n\text{-}1)}\left(\hat{\sigma}_{e_1} / \sqrt{n}\right)\right] \qquad (3)$$

where $t_{\alpha/2;(n-1)}$ is the upper $\alpha/2$ quantile of Student's *t* distribution with $(n-1)$ *df*. After rearrangement, eq. 3 is

$$P\left(overlap\right) = \left[\left|\sqrt{n}\left(\bar{Y}_{1.} \text{-} \bar{Y}_{2.}\right)\right| < t_{\alpha/2;(n\text{-}1)}\left(\hat{\sigma}_{e_1} + \hat{\sigma}_{e_2}\right)\right] \qquad (4)$$

One of the terms in eq. 4 is $\left(\bar{Y}_{1.} - \bar{Y}_{2.}\right)$. When data are collected in an unpaired setting (eq. 1), the standard error of $\left(\bar{Y}_{1.} - \bar{Y}_{2.}\right)$ is estimated by $\sqrt{\left(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_{e_2}^2\right)/n}$. Thus, if we divide eq. 4 by $\sqrt{\left(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_{e_2}^2\right)}$, and then recall that $t_{\alpha/2;n-1}^2 = F_{\alpha;1,n-1}$, eq. 4 can be written as

$$P\left(\text{overlap}\right) = P\left[\left(\frac{\left(\bar{Y}_{1.} \text{-} \bar{Y}_{2.}\right)}{\sqrt{\left(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_{e_2}^2\right)/n}}\right)^2 < F_{\alpha;1,n\text{-}1} \frac{\left(\hat{\sigma}_{e_1} + \hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}_{e_1}^2 + \hat{\sigma}_{e_2}^2}\right] \qquad (5)$$

Replacing the tabular *t* (eq. 4) or *F* (eq. 5) with the value of '1' yields the probability of overlapping standard errors. The large sample probability that standard errors around two means overlap is

$$P\left(\text{overlap}\right) = 1 \text{-} \left\{2\left(1 \text{-} \Phi\left[\frac{\sigma_{e_1} + \sigma_{e_2}}{\sqrt{\sigma_{e_1}^2 + \sigma_{e_2}^2}}\right]\right)\right\} \qquad (6)$$

where $\Phi$ is the cumulative standard normal distribution. The large sample probability that two $\left(1-\alpha\right)100\%$ confidence intervals overlap (*cf* Goldstein & Healy[31]) is obtained by multiplying the term in square brackets in eq. 6 by $z_{\alpha/2}$, the upper $\alpha/2$ quantile of the standard normal distribution.

Afshartous & Preston[32] studied this problem in situations with 'dependent data' (i.e., paired samples); their approach, however, was limited to cases involving a fixed effects model. In paired settings, it is usually the case that the basis for pairing represents a random nuisance variable, and so a mixed model (eq. 2) is generally more appropriate in most applications. To extend these considerations to a paired setting with a random nuisance effect, we note that the standard error of a treatment mean, $\bar{Y}_{i.}$, is $\sqrt{\left(\sigma_{e_i}^2 + \sigma_b^2\right)/n}$, where $\sigma_{e_i}^2 + \sigma_b^2$ is estimated by $\sum_j^n \left(\bar{Y}_{ij} - \bar{Y}_{i.}\right)^2 / \left(n-1\right)$. Thus, eq. 3 becomes

$$P\left(\text{overlap}\right) = 1 - P\left[\bar{Y}_{1.} + t_{\alpha/2;(n-1)}\sqrt{\left(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_b^2\right)/n} < \bar{Y}_{2.} - t_{\alpha/2;(n-1)}\sqrt{\left(\hat{\sigma}_{e_2}^2 + \hat{\sigma}_b^2\right)/n}\right] -$$

$$P\left[\bar{Y}_{2.} + t_{\alpha/2;(n\text{-}1)}\sqrt{\left(\hat{\sigma}_{e_2}^2 + \hat{\sigma}_b^2\right)/n} < \bar{Y}_{1.} \text{-} t_{\alpha/2;(n\text{-}1)}\sqrt{\left(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_b^2\right)/n}\right] \qquad (7)$$

**Citation:** Wester DB. Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis. *Biom Biostat Int J.* 2018;7(1):73‒85. DOI: 10.15406/bbij.2018.07.00192

and eq. 4 becomes

$$P\left(\text{overlap}\right)=\left[\left|\sqrt{n}\left(\bar{Y}_{1.}-\bar{Y}_{2.}\right)\right|<t_{\alpha/2;(n-1)}\left(\sqrt{\hat{\sigma}_{e_1}^2+\hat{\sigma}_b^2}+\sqrt{\hat{\sigma}_{e_2}^2+\hat{\sigma}_b^2}\right)\right] \quad (8)$$

In a paired setting (eq. 2), the standard error of $\bar{Y}_{1.}-\bar{Y}_{2.}$ is estimated by $\sqrt{\left(\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2-2\hat{\sigma}_{e_{12}}\right)/n}$ . Thus, we divide eq. 8 by $\sqrt{\left(\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2-2\hat{\sigma}_{e_{12}}\right)}$ and eq. 5 becomes

$$P\left(\text{overlap}\right)=P\left[\left(\frac{\left(\bar{Y}_{1.}\text{-}\bar{Y}_{2.}\right)}{\sqrt{\left(\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2\text{-}2\hat{\sigma}_{e_{12}}\right)/n}}\right)^2<F_{\alpha;1,n\text{-}1}\frac{\left(\sqrt{\hat{\sigma}_{e_1}^2+\hat{\sigma}_b^2}+\sqrt{\hat{\sigma}_{e_2}^2+\hat{\sigma}_b^2}\right)^2}{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2\text{-}2\hat{\sigma}_{e_{12}}}\right] \quad (9)$$

The large sample probability that standard errors around two means overlap in a paired setting is

$$P\left(\text{overlap}\right)=1-\left\{2\left(1\text{-}\Phi\left[\frac{\sqrt{\sigma_{e_1}^2+\sigma_b^2}+\sqrt{\sigma_{e_2}^2+\sigma_b^2}}{\sigma_{e_1}^2+\sigma_{e_2}^2\text{-}2\sigma_{e_{12}}}\right]\right)\right\} \quad (10)$$

where $\Phi$ is the cumulative standard normal distribution. Multiplying the term in square brackets in eq. 10 by $z_{\alpha/2}$ gives the large—sample probability that two $\left(1-\alpha\right)100\%$ confidence intervals overlap.

## Monte carlo results: unpaired *t* test (CRD)

Each time in the Monte Carlo simulation, when two confidence intervals overlapped it was also true that $F_c=\left(\sqrt{n\left(\bar{Y}_{1.}-\bar{Y}_{2.}\right)}/\sqrt{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2}\right)^2$ was less than $F_{\alpha;1,n-1}\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2}$ (Table 1), providing empirical support for the simulations. A similar conclusion applies to incidences of overlapping standard errors and $F_c<\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2}$ (eq. 5 with the tabular *F* replaced with the value of '1').

When variances are equal, the term $\frac{\sigma_{e_1}+\sigma_{e_2}}{\sqrt{\sigma_{e_1}^2+\sigma_{e_2}^2}}$ is equal to $\sqrt{2}$ and the probability that standard errors overlap (eq. 6) reaches its maximum value of *P* = 0.8427 (corresponding with a type I error rate of about 16%), illustrated in Table 1 with increasing sample size. With unequal variances, the probability that standard errors overlap decreases and the type I error rate increases; Table 1 illustrates this when $\sigma_{e_2}/\sigma_{e_1}=4$ , in which case the probability of overlap decreases to ~77.58% (and the type I error increases to about 23%) with increasing sample size. As the difference between variances increases, $\frac{\sigma_{e_1}+\sigma_{e_2}}{\sqrt{\sigma_{e_1}^2+\sigma_{e_2}^2}}$ approaches 1, and the probability that standard errors overlap approaches a value of 68.27% (corresponding to a type I error rate of ~32%; (Figure 2). In contrast to results based on overlapping standard errors, the frequency of overlapping 95% confidence intervals approached 99.44% (0.56% type I error rate) with homogeneous variances and 98.25% (1.75% type I error rate) when $\sigma_{e_2}/\sigma_{e_1}=4$ (Table 1). With increasing heteroscedasticity, the probability that two confidence intervals overlap approaches 95% (a type I error rate of 5%) with large sample sizes.

## Monte carlo results: paired *t* test (RBD)

Overlapping confidence intervals in the Monte Carlo simulation in a paired setting coincided with $F_c=\left(\frac{\sqrt{n}\left(\bar{Y}_{1.}-\bar{Y}_{2.}\right)}{\sqrt{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2-2\hat{\sigma}_{e_{12}}}}\right)^2$ being less than $F_{\alpha;1,n-1}\frac{\left(\sqrt{\hat{\sigma}_{e_1}^2+\hat{\sigma}_b^2}+\sqrt{\hat{\sigma}_{e_2}^2+\hat{\sigma}_b^2}\right)^2}{\hat{\sigma}_{e_1}^2+\hat{\sigma}_{e_2}^2-2\hat{\sigma}_{e_{12}}}$ ; a similar conclusion applies to the overlap of standard errors (Table 2).

In the expression for the probability that two standard errors overlap for paired sampling (eq. 10), the term $\frac{\sqrt{\sigma_{e_1}^2+\sigma_b^2}+\sqrt{\sigma_{e_2}^2+\sigma_b^2}}{\sqrt{\sigma_{e_1}^2+\sigma_{e_2}^2\text{-}2\sigma_{e_{12}}}}$ reaches a minimum value equal to 1 when $\sigma_{e_1}^2=\sigma_{e_2}^2=\sigma_e^2,\sigma_b^2=0$ , and the covariance, $\sigma_{e_{12}}$ , is negative but such that $\left|\sigma_{e_{12}}\right|\to\sigma_e^{2-}$ (the latter condition to ensure that $\sum$, the $tXt$ variance—covariance matrix, is positive definite), and under these conditions the probability that two standard errors overlap reaches its minimum value of approximately 68% (corresponding to a type I error rate of about 32%; (Figure 3). With homogeneous variances between treatments and $\sigma_b^2=0$ , the probability that standard errors overlap increases to its maximum value of 100% (corresponding to the 0% probability of a type I error) when the covariance between treatments is positive and $\sigma_{e_{12}}\to\sigma_e^{2-}$ .

The term $\frac{\sqrt{\sigma_{e_1}^2+\sigma_b^2}+\sqrt{\sigma_{e_2}^2+\sigma_b^2}}{\sqrt{\sigma_{e_1}^2+\sigma_{e_2}^2\text{-}2\sigma_{e_{12}}}}$ in eq. 10 also increases beyond its minimum value when variances within treatments are heterogeneous ( $\sigma_{e_1}^2\neq\sigma_{e_2}^2$ ) and/or $\sigma_b^2>0$ and/or $\sigma_{e_{12}}\neq0$ (with $\sum$ positive definite), under which conditions the probability that standard errors overlap increases and can approach 1. These general conditions are illustrated in Table 2 for 3 specific combinations of variances within treatments and covariances between treatments as well as 3 values for the block variance. Figure 3 illustrates a scenario where the probability of overlapping standard errors extends across its range of ~68% to ~100%—when variances within treatments are homogeneous, the block variance equals 0, and the covariance between treatments, $\sigma_{e_{12}}$ , varies—as well as selected cases when $\sigma_{e_i}^2$ are unequal and $\sigma_b^2>0$ . As with the unpaired setting, the type I error rate for the paired *t* test is approximately 5% regardless of sample size, the variances and covariances of errors, or the variance of the block effect.

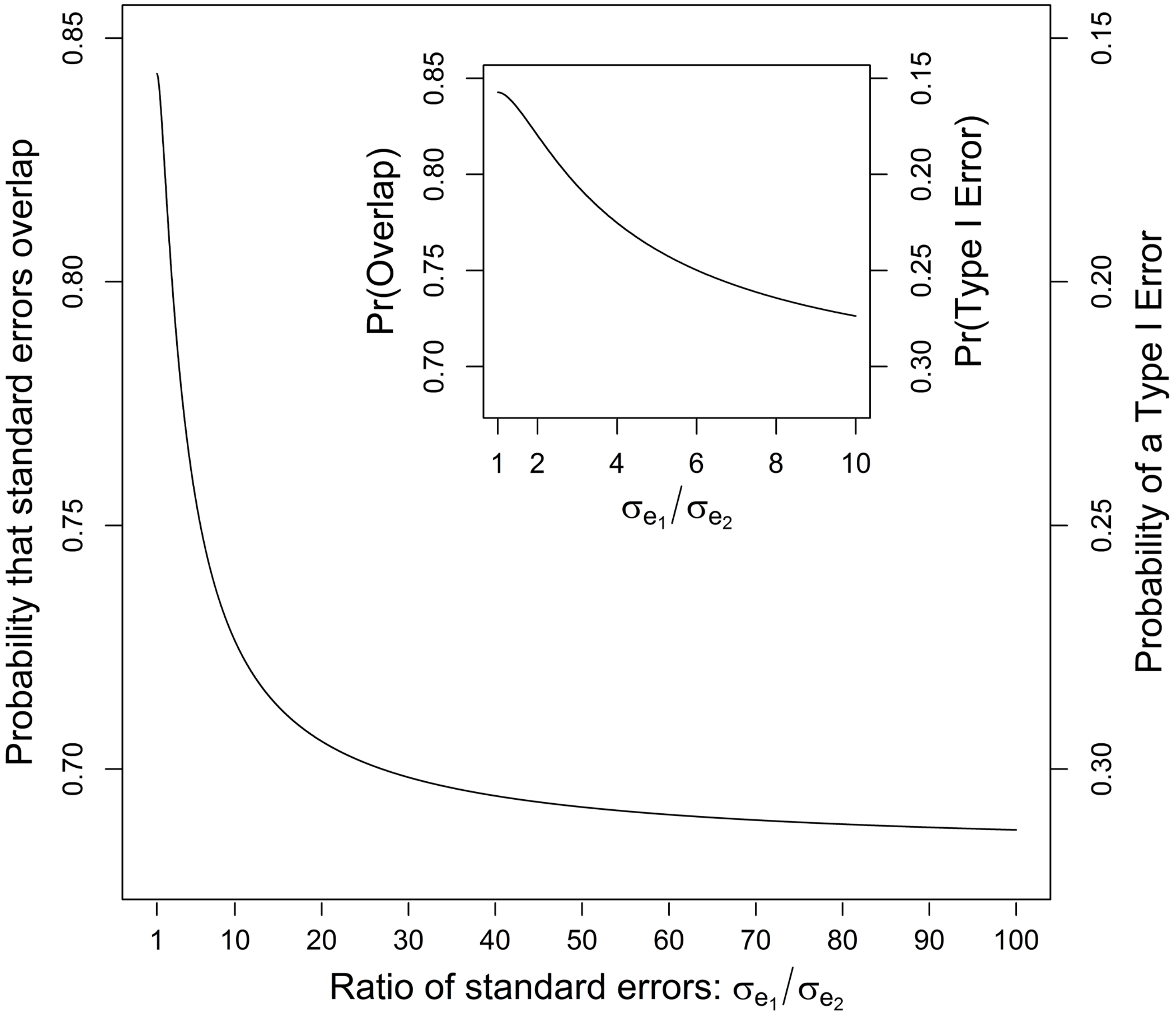

**Figure 2** Relationship between the probability that standard errors around two treatment means (when $\mu_1 = \mu_2$) collected in an unpaired setting overlap (as well as the associated type I error rate) as a function of the ratio of the standard errors of each mean, $\sigma_{e_1} / \sigma_{e_2}$, ranging from 1 to 100. Inserted figure: $\sigma_{e_1} / \sigma_{e_2}$, ranges from 1 to 10.

## Regarding one—sided tests

A confidence interval is simply a 'scaled' standard error. Meehl et al.[33] argued that non—overlapping scaled standard error bars, $\overline{Y}_{i.} \pm t_{\alpha;2n-2} S_i$, where $S_i = \sqrt{\Sigma\left(\overline{Y}_{ij} - \overline{Y}_{i.}\right)^2 / (n-1)}$, is equivalent to a one—sided $\alpha$ —level $t$ test of the difference between two means. It should be noted that the *df* of their tabular Student's $t$ statistic correspond with the error term in an unpaired setting (and are therefore different from the *df* in eq. 3); additionally, $S_i$, as defined by Meehl et al.[33] is a standard deviation rather than a standard error. That is, they suggested that $\left(\overline{Y}_{1.} + t_{\alpha;2n-2} S_1\right) < \left(\overline{Y}_{2.} - t_{\alpha;2n-2} S_2\right)$ is equivalent to rejecting $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 < \mu_2$ with a one—sided $\alpha$

—level *t* test. This is incorrect. Even when a standard error, $S_i/\sqrt{n}$, is used, the asymptotic probability that two such scaled standard error bars overlap is

$$P\left(\text{overlap}\right)=\Phi\left[z_\alpha\frac{\sigma_{e_1}+\sigma_{e_2}}{\sqrt{\sigma^2_{e_1}+\sigma^2_{e_2}}}\right] \qquad (11)$$

Thus, with $\alpha=0.05$, this probability approaches $1-\alpha=0.95$ only when $\sigma^2_{e_i}$ are very different (and $\frac{\sigma_{e_1}+\sigma_{e_2}}{\sqrt{\sigma^2_{e_1}+\sigma^2_{e_2}}}$ approaches 1); and this probability approaches 0.99 when variances between treatments are equal, leading to a type I error rate ~1% rather than ~5%.

**Table 1** Summary of *N* = 10,000 Monte Carlo simulations of a completely randomized design with two treatments (with $\mu_1=\mu_2$) (unpaired *t* test, $\alpha=0.05$). Values in the table are the relative frequency of overlapping standard errors or overlapping 95% confidence intervals and the relative frequency that $F_c=\left(\sqrt{n}\left(\bar{Y}_{1.}-\bar{Y}_{2.}\right)/\sqrt{\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}}\right)^2$ or $t_c=|\left(\bar{Y}_{1.}-\bar{Y}_{2.}\right)/\sqrt{\left(\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}\right)/n}$ | were less than indicated critical values with $v=t\left(n-1\right)$ or $v'=\frac{\left(\hat{\sigma}^2_{e_1}/n+\hat{\sigma}^2_{e_2}/n\right)^2}{\left(\hat{\sigma}^2_{e_1}/n\right)^2/\left(n-1\right)+\left(\hat{\sigma}^2_{e_2}/n\right)^2/\left(n-1\right)}$ degrees of freedom. Asymptotic values are from eq. 6. Type I error rates are 1 – tabulated values

| | **Standard errors** | **Confidence intervals** | **t Test** |
|---|---|---|---|
| **Sample size (n)** | $\text{Overlap}=F_c<\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}}$ | $\text{Overlap}=F_c<F_{\alpha/2,v}\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}}$ | $\lvert t_c\rvert<t_{\alpha/2;v}$ |
| | | Homogeneous variances:<br>$\sigma_{e_1}=4;\sigma_{e_2}=4$ | |
| 5 | 0.7847 | 0.9952 | 0.9503 |
| 10 | 0.8161 | 0.9936 | 0.9468 |
| 20 | 0.8310 | 0.9941 | 0.9503 |
| 30 | 0.8327 | 0.9941 | 0.9478 |
| 50 | 0.8369 | 0.9949 | 0.9523 |
| 100 | 0.8399 | 0.9952 | 0.9513 |
| 1000 | 0.8492 | 0.9950 | 0.9491 |
| Asymp | 0.8427 | 0.9944 | 0.9500 |
| | **Standard Errors:** | **Confidence Intervals:** | **t Test:** |
| **Sample Size (n)** | $\text{Overlap}=F_c<\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}}$ | $\text{Overlap}=F_c<F_{\alpha/2,v}\frac{\left(\hat{\sigma}_{e_1}+\hat{\sigma}_{e_2}\right)^2}{\hat{\sigma}^2_{e_1}+\hat{\sigma}^2_{e_2}}$<br>Heterogeneous variances:<br>$\sigma_{e_1}=4;\sigma_{e_2}=16$ | $\lvert t_c\rvert<t_{\alpha/2;v'}$ |
| 5 | 0.7146 | 0.9813 | 0.9456 |
| 10 | 0.7424 | 0.9841 | 0.9471 |
| 20 | 0.7603 | 0.9815 | 0.9482 |
| 30 | 0.7678 | 0.9827 | 0.9510 |
| 50 | 0.7627 | 0.9849 | 0.9496 |
| 100 | 0.7688 | 0.9836 | 0.9486 |
| 1000 | 0.7758 | 0.9841 | 0.9523 |
| Asymp | 0.7748 | 0.9825 | 0.9500 |

## Concerning experiment—wise and comparison—wise error rates

The foregoing dealt with experiments that involved two treatments. When an experiment involves more than two treatments and the overall research objectives include all possible pairwise comparisons among treatments, the investigator should be concerned with the distinction between experiment—wise and comparison—wise error rate. There are many available tests.

A commonly—used analytical strategy, the protected (or Fisher's) Least Significant Different test (FLSD), involves an initial $\alpha$ —level *F* test of overall treatment mean equality. When this *F* test is not significant then pairwise comparisons are not made; when the initial *F* test is significant, then all possible pairwise comparisons are made with an $\alpha$ —level *t* test. The initial *F* test controls the experiment—wise error rate at $\alpha$ and the comparison—wise error rate is less than $\alpha$ when the initial *F* test is used to guide whether or not pairwise comparisons are performed. If, however, all possible pairwise comparisons are made with an $\alpha$ —level *t* test without first performing the initial *F* test (LSD), then although the comparison—wise error rate = $\alpha$ the experiment—wise error rate > $\alpha$ (and increases as the number of treatments increase).

To illustrate these principles, a completely randomized design with $t = 5$ treatments, $n = 30$ replications and homogeneous variances of experimental errors was used with $N = 10{,}000$ Monte Carlo simulations in which population treatment means were equal. The initial *F* test was significant in 511 of the 10,000 experiments, confirming an experiment—wise error rate of ~5%. When all possible pairwise comparisons were made with an $\alpha$ —test *t* test (LSD, without consulting the initial *F* test), 2,860 of the 10,000 experiments included at least one type 1 error (experiment—wise error rate = 28.6%) and the comparison—wise error rate was 5.097%. However, if all possible pairwise comparisons were made with an $\alpha$ —level *t* test only when the initial *F* test of overall treatment mean equality was significant (FLSD), then the comparison—wise error rate was 1.615%. These results are similar to those reported in Table 3 of Carmer & Swanson:[34] with 4,000 experiments and 5 treatments, their experiment—wise error rate of the protected LSD test was 4.8% and the comparison error rate was 1.82%; when unprotected $\alpha$ —level *t* tests were used the experiment—wise error was 25.6% and the comparison—wise error rate was 4.99%.

How are these principles affected when treatments are compared, not with $\alpha$ —level *t* tests, but with overlapping standard errors or overlapping $(1-\alpha)$ 100% confidence intervals? In the 10,000–experiment Monte Carlo simulation described above, when the initial *F* test was not consulted, then 82.953% and 99.467% of comparisons had overlapping standard errors and overlapping $(1-\alpha)$100% confidence intervals, respectively; these values are similar to those reported in Table 1 (and correspond to 17.047% and 0.533% comparison—wise error rates, respectively). Furthermore, the experiment—wise error rate was 100% for both overlapping standard errors and overlapping confidence intervals. If pairwise comparisons among treatments were made only if the initial *F* test was significant, then 2.529% and 4.713% of comparisons had overlapping standard errors and overlapping $(1-\alpha)$100% confidence intervals, respectively (corresponding to 97.471% and 95.287% comparison—wise error rates, respectively).

## Discussion

In many applied sciences, most researchers consider it desirable to determine the significance level of their tests of hypotheses, and although it is widely accepted that there is nothing sacrosanct about (say) a 5%—level test, it is still common in many fields to adopt $\alpha = 0.05$ or $\alpha = 0.10$ when drawing conclusions and advancing recommendations. Tables 1 & 2, Figures 2 & 3 shows that basing inferences on overlapping standard errors or confidence intervals around treatment means leaves much to be desired: this practice does not control type I error rates at commonly—accepted levels.

For a CRD with two treatments, using overlapping standard errors to compare treatment means is a decision tool with a type I error rate that ranges from (at best) 16% when variances are homogeneous and sample sizes are large to nearly 32% when variances are extremely heterogeneous. For a CRD with five treatments (and homogeneous variances), comparison—wise type I error rates based on overlapping standard errors (17.05%) were similar to results observed in two—treatment experiments. However, when pairwise comparisons were performed only after a significant *F* test of overall treatment mean equality, then comparison—wise type I error rates were 97.47%. Clearly, inspection alone of 'overlapping standard errors' is not a reliable substitute for a formal $\alpha$ —level hypothesis test (*cf* Cherry[35]). In contrast, using overlapping 95% confidence intervals to compare treatment means is a much more conservative strategy in two—treatment experiments, with a type I error rate as low as 0.56% when variances are homogeneous to (at most) 5% when variances are extremely heterogeneous. This procedure was similarly conservative in five treatment experiments (with homogeneous variances) when unprotected $\alpha$ —level *t* tests were used. However, if $\alpha$ —level *t* tests are used only following a significant overall *F* test, then comparison—wise type I error rates based on 95% confidence intervals are high (95.29%).

Similar considerations apply to a paired setting, although there are more factors that affect the probability that standard errors (or confidence intervals) overlap. For example, a negative covariance between treatments reduces the probability of overlap and a positive covariance increases the probability of overlap (relative to the independence case). And increasing variability attributable to the random block effect increases probability of overlap. Thus, in an RBD with $t > 2$ treatments, where it is likely that the correlation between pairs of treatments will vary (even when sphericity is satisfied), the probability of overlap will vary as the correlation between pairs of treatments varies.

If one wants to compare means in a two—treatment experiment and have an idea of the type I error rate for the comparison, then using overlapping standard errors gives a very liberal idea of that error rate (on average, more differences will be declared than actually exist) and using overlapping confidence intervals gives a very conservative idea of that error rate (i.e., the test will have lower power than the nominal level). Similar conclusions apply to a five—treatment experiment with inferences based on treatment mean comparisons that are not protected by a significant *F* test on overall treatment mean equality. If, however, treatment mean comparisons are made only following a significant *F* test in a five—treatment experiment, comparison—wise type I error rates are high (~95%) whether overlapping standard errors are used or overlapping 95% confidence intervals are used. Put another way, neither approach—using overlapping standard errors or using overlapping confidence intervals—gets you where you want to be if the goal is to attach an accurate *P*—value to an inference, a conclusion reached by Schenker & Gentleman[36] as well.

In fact, the only way to compare treatment means with a type I error rate that approximates the nominal level is to formally test

$H_0:\mu_1=\mu_2$ with Student's *t* test which controls the error rate for all sample sizes as long as assumptions are satisfied. And this formal test is equivalent to estimating a confidence interval around the difference between treatment means—if this interval includes zero, then there is no significant difference between treatments. In fact, the most complete analysis and presentation of results would include

i. Treatment means and standard errors for these means, with a *t* test of the difference between means (following a significant *F* test in experiments with more than two treatments), its *df* and *P*—value (Figure 1d); and

ii. A confidence interval around this difference. With these results, the investigator has an estimate both of significance and effect size, with the added value of an interval estimate: a confidence interval around a difference might include zero but its width may be interpreted differently by subject—matter specialists.

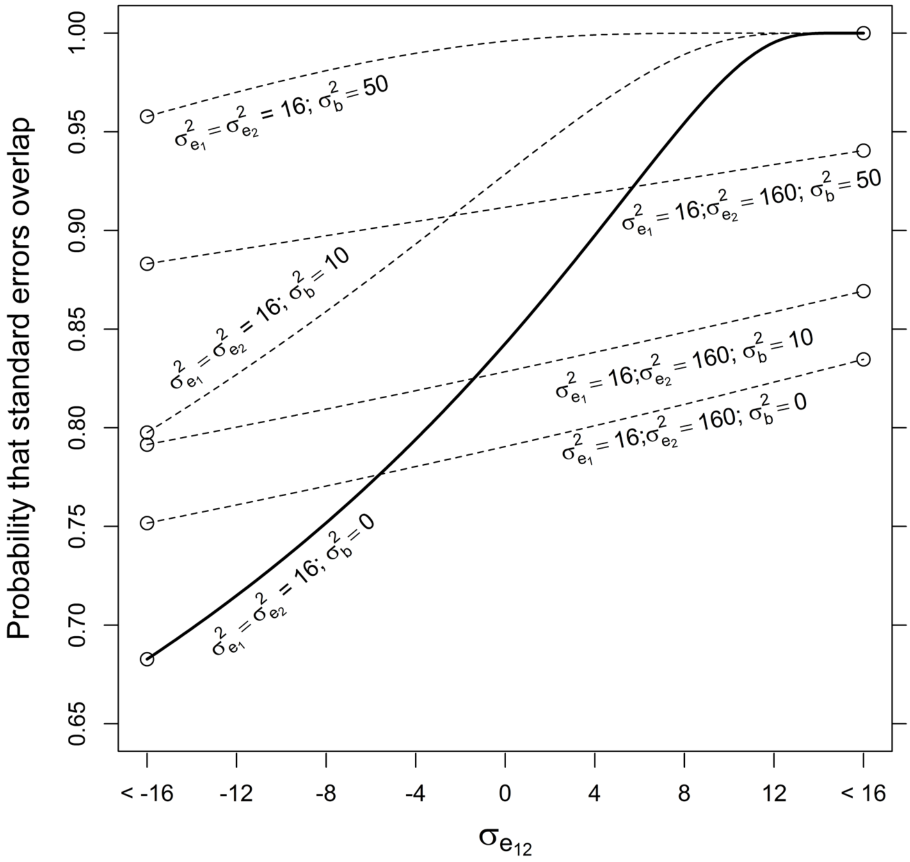

**Figure 3** Relationship between the probability that standard errors around two treatment means (when $\mu_1=\mu_2$) collected in a paired setting overlap (as well as the associated type I error rate) as a function of the covariance between treatments for several combinations of within–treatment variances, $\sigma^2_{e_i}$, and block variance, $\sigma^2_b$.

This recommendation in no way discourages the presentation of standard errors and/or confidence intervals around treatment means (e.g., Figure 1b & 1c): this is vital descriptive information which enables an assessment of the precision of our estimates of treatment means. But standard errors (or confidence intervals) around each of two means—and whether they overlap or not—is not a reliable indicator of the statistical significance associated with a test of the hypothesis of mean equality, despite common interpretations to the contrary: for this purpose, interest must shift to a confidence interval around the difference between two means.

One might argue that the last column in Table 1 and the last 3 columns in Table 2 are not necessary. The distribution of the *t* statistic is known, and so the nominal $\alpha$ level is exact: these columns simply reflect simulation variability, and tell us nothing about the behavior of the test statistic that we do not already know. And this is absolutely true. However, a common criticism of—and misunderstanding about—null hypothesis testing is expressed in the following claim: 'Give me a large enough sample size and I will show you a highly significant difference between two treatments whether a true difference exists or not' (a claim once made to me by a journal editor). These columns clearly show that this is not correct: type I error rates of the *t* test were about 5% whether the sample size was *n* = 5 or *n* = 1,000: strictly speaking, the type I error rate of a *t* test of hypothesis is not a function of sample size (nor is it a function of the variance of the random

**Citation:** Wester DB. Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis. *Biom Biostat Int J.* 2018;7(1):73–85. DOI: 10.15406/bbij.2018.07.00192

nuisance variable in a paired setting). Statistical power *is*, of course, affected by sample size; and whereas the two—fold criticism that (1) the 'null effect' is likely never precisely zero (but see Berger & Delampady[37], who argued that the *P*—value corresponding to a point null hypothesis is a reasonable response to this criticism), and (2) even a small difference may be significant with a large enough sample size (e.g., Lin et al.[38]) is well—founded, this is not the issue addressed in this paper. Furthermore, the probability that two standard errors overlap (when the null hypothesis is true) is only a weak function of sample size: this probability was ~0.83—similar to the asymptotic probability of ~0.84—with sample sizes as small as *n* = 20. Likewise, the probability that two confidence intervals overlap is not (much of) a function of sample size: this probability was very similar to the asymptotic probability with sample sizes throughout the range of *n* = 5 to *n* = 1, 000.

**Table 2** Summary of *N* = 10,000 Monte Carlo simulations of randomized block design with two treatments (with $\mu_1 = \mu_2$) (paired *t* test, $\alpha = 0.05$). Values in the table are the relative frequency of overlapping standard errors or overlapping 95% confidence intervals and the relative frequency that $F_c = \left(\sqrt{n}\left(\bar{Y}_{1.} - \bar{Y}_{2.}\right) / \sqrt{\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_{e_2} - 2\hat{\sigma}_{e_{12}}}\right)^2$ or $t_c = |\left(\bar{Y}_{1.} - \bar{Y}_{2.}\right) / \sqrt{\left(\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_{e_2} - 2\hat{\sigma}_{e_{12}}\right)/n}$ were less than indicated critical values with $v = (t-1)(n-1)$. Asymptotic values are from eq. 10. Type I error rates are 1 – tabulated values

| Sample size (*n*) | Standard errors: | | | Confidence intervals: | | | *t* test | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\text{Overlap} = F_c < \frac{\left(\sqrt{\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_b} + \sqrt{\hat{\sigma}^2_{e_2} + \hat{\sigma}^2_b}\right)^2}{\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_{e_2} - 2\hat{\sigma}_{e_{12}}}$ | | | $\text{Overlap} = F_c < F_{\alpha/2,v} \frac{\left(\sqrt{\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_b} + \sqrt{\hat{\sigma}^2_{e_2} + \hat{\sigma}^2_b}\right)^2}{\hat{\sigma}^2_{e_1} + \hat{\sigma}^2_{e_2} - 2\hat{\sigma}_{e_{12}}}$ | | | $\lvert t_c \rvert < t_{\alpha/2;v}$ | | |
| | $\sigma^2_b$ | | | $\sigma^2_b$ | | | $\sigma^2_b$ | | |
| | 0 | 4 | 16 | 0 | 4 | 16 | 0 | 4 | 16 |
| Variance-covariance structure: | | | | | | | | | |
| 5 | 0.7966 | 0.8386 | 0.9047 | 0.9949 | 0.9976 | 0.9987 | 0.9506 | 0.9506 | 0.9506 |
| 10 | 0.8189 | 0.8621 | 0.9320 | 0.9952 | 0.9982 | 0.9996 | 0.9513 | 0.9513 | 0.9513 |
| 20 | 0.8291 | 0.8732 | 0.9429 | 0.9951 | 0.9977 | 1.0000 | 0.9455 | 0.9455 | 0.9455 |
| 30 | 0.8300 | 0.8749 | 0.9428 | 0.9944 | 0.9986 | 1.0000 | 0.9454 | 0.9454 | 0.9454 |
| 50 | 0.8321 | 0.8784 | 0.9465 | 0.9932 | 0.9972 | 0.9999 | 0.9485 | 0.9485 | 0.9485 |
| 100 | 0.8442 | 0.8857 | 0.9556 | 0.9949 | 0.9985 | 0.9999 | 0.9526 | 0.9526 | 0.9526 |
| 1000 | 0.8431 | 0.8855 | 0.9541 | 0.9946 | 0.9982 | 0.9999 | 0.9498 | 0.9498 | 0.9498 |
| Asymp | 0.8427 | 0.8862 | 0.9545 | 0.9944 | 0.9981 | 0.9999 | 0.9500 | 0.9500 | 0.9500 |
| Variance-covariance structure: $\sigma_{e_1} = 4; \sigma_{e_2} = 4; \sigma_{e_{12}} = 4$ | | | | | | | | | |
| 5 | 0.8487 | 0.8818 | 0.9361 | 0.9981 | 0.9990 | 0.9995 | 0.9521 | 0.9521 | 0.9521 |
| 10 | 0.8738 | 0.9076 | 0.9596 | 0.9986 | 0.9993 | 0.9999 | 0.9519 | 0.9519 | 0.9519 |
| 20 | 0.8852 | 0.9214 | 0.9678 | 0.9984 | 0.9994 | 1.0000 | 0.9472 | 0.9472 | 0.9472 |
| 30 | 0.8830 | 0.9227 | 0.9722 | 0.9987 | 0.9996 | 1.0000 | 0.9476 | 0.9476 | 0.9476 |
| 50 | 0.8900 | 0.9245 | 0.9736 | 0.9978 | 0.9995 | 1.0000 | 0.9478 | 0.9478 | 0.9478 |
| 100 | 0.8949 | 0.9345 | 0.9800 | 0.9988 | 0.9993 | 1.0000 | 0.9533 | 0.9533 | 0.9533 |
| 1000 | 0.8962 | 0.9318 | 0.9807 | 0.9988 | 0.9999 | 0.9999 | 0.9508 | 0.9508 | 0.9508 |
| Asymp | 0.8975 | 0.9321 | 0.9791 | 0.9986 | 0.9997 | 0.9999 | 0.9500 | 0.9500 | 0.9500 |
| Variance-covariance structure: $\sigma_{e_1} = 4; \sigma_{e_2} = 4; \sigma_{e_{12}} = -4$ | | | | | | | | | |
| 5 | 0.7467 | 0.7987 | 0.8780 | 0.9914 | 0.9954 | 0.9977 | 0.9501 | 0.9501 | 0.9501 |
| 10 | 0.7710 | 0.8196 | 0.9047 | 0.9995 | 0.9938 | 0.9990 | 0.9502 | 0.9502 | 0.9502 |
| 20 | 0.7767 | 0.8290 | 0.9108 | 0.9881 | 0.9950 | 0.9991 | 0.9465 | 0.9465 | 0.9465 |
| 30 | 0.7780 | 0.8277 | 0.9127 | 0.9869 | 0.9958 | 0.9995 | 0.9456 | 0.9456 | 0.9456 |
| 50 | 0.7807 | 0.8312 | 0.9163 | 0.9855 | 0.9931 | 0.9997 | 0.9486 | 0.9486 | 0.9486 |
| 100 | 0.7907 | 0.8439 | 0.9263 | 0.9876 | 0.9949 | 0.9993 | 0.9518 | 0.9518 | 0.9518 |
| 1000 | 0.7970 | 0.8436 | 0.9260 | 0.9870 | 0.9944 | 0.9998 | 0.9497 | 0.9497 | 0.9497 |
| Asymp | 0.7941 | 0.8427 | 0.9264 | 0.9868 | 0.9944 | 0.9996 | 0.9500 | 0.9500 | 0.9500 |

The basic thesis of this paper is that using 'overlap'—either of standard errors or confidence intervals estimated around each treatment mean—is unwise if the goal is to draw an inference about the equality of treatments means at commonly—used significance levels. In the area of multivariate analysis of variance, however[39,41], developed an analytical/graphical comparison between treatments that combines the intuitive appeal of an easy—to—understand display of overlap—in this case, between 'error' and 'treatment' ellipses—with the inferential guarantee of a specified $\alpha$ —level test of hypothesis. To appreciate this approach, it is helpful to recall that a confidence interval around a treatment mean is a unidimensional concept: $\bar{Y}_{1.} \pm t_{\alpha/2;(n-1)}\left(\hat{\sigma}_{e_1} / \sqrt{n}\right)$ defines a region in a one—dimensional space such that there is a $(1-\alpha)$ 100% chance that the interval includes the true population mean. When two (or more) dependent variables are measured, then a confidence ellipse (or ellipsoid) can be formed that defines a region in 2— (or higher—) dimensional space that has a $(1-\alpha)$ 100% chance of encompassing the true population centroid. Friendly and Fox et al.[39-41] applied these ideas to 'error' and 'treatment' ellipses based on the sums of squares and cross—products matrices for the treatment and error effects in a multivariate analysis of variance. In this multivariate analysis, the treatment ellipse (or line) can be scaled so that, if it is included within the boundary of the error ellipse in all dimensions in the space defined by dependent variables (i.e., if the treatment ellipse 'does not overlap' the error ellipse), then there is no statistical difference (at the specified $\alpha$ level) between the treatments (Figure 4a). If the treatment ellipse extends beyond the boundaries of the error ellipse (i.e., 'overlaps' the error ellipse), then centroids differ between treatments in the dimension defined by the linear combination of dependent variables that is represented by the major axis of the ellipse (corresponding to Roy's Maximum Root criterion with test size $\alpha$ ) (Figure 4b).

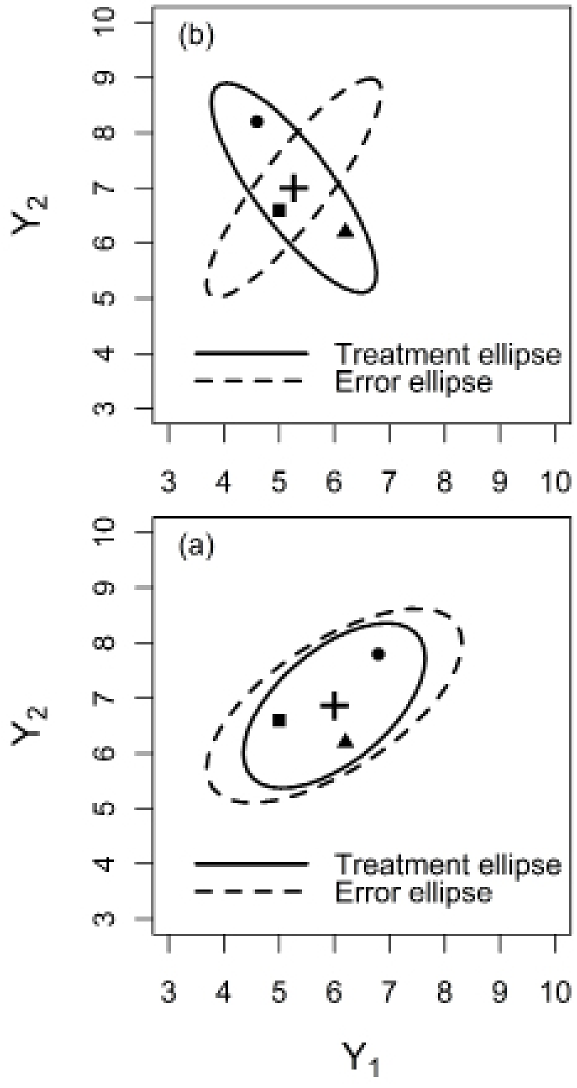

**Figure 4** (a) A non–significant ($P = 0.07$) multivariate F test (Roy's Maximum Root Criterion) for a case with three treatments and two dependent variables, $Y_1$ and $Y_2$; therefore, the treatment ellipse lies inside the error ellipse, indicating that there is no linear combination of $Y_1$ and $Y_2$ along which treatments are significantly different (univariate tests for both dependent variables are not significant at $\alpha > 0.10$). (b) A significant ($P < 0.0001$) multivariate F test (Roy's Maximum Root Criterion); therefore, the treatment ellipse extends outside (i.e., overlaps) the boundaries of the error ellipse in the direction of the linear combination of $Y_1$ and $Y_2$ that maximizes a difference between treatment centroids that is statistically significant (univariate tests for both dependent variables are not significant at $P = 0.06$). Solid square, round and triangular dots represent treatment centroids; the 'X' is the overall centroid.

## Summary

The proper interpretation of standard error bars and confidence intervals has received attention in the literature (e.g., Goldstein & Healy[31]; Payton et al.;[26] Payton et al.;[27] Cummings et al.;[42] Cummings & Finch[43]; Afshartous & Preston[32]). It is also true, however, that misinterpretation is common, with the possibility of misleading or erroneous inferences about statistical differences. All of the quotations cited in the introduction (above) are from refereed journals in disciplines as diverse as wildlife science, plant ecology, materials science, facial recognition, psychology, and climate science—and so these papers were reviewed by peers as well as associate editors and editors prior to publication. Schenker & Gentleman[36] identified over 60 articles in 22 different journals in the medical field which either used or recommended the method of overlap to evaluate inferential significance. Clearly, this approach is widespread and many applied scientists believe the practice is sound. The results in this paper indicate otherwise. And whereas there are many reasons why this practice should be avoided, there is one reason that stands out: when an author declares 'To more generally compare differences [among treatments], we define non—overlapping standard error values as representative of significantly different [survival estimates]',[44] the effect of such arbitrary redefinition will surely lead to confusion: what one investigator has defined as 'significant' will mean something very different from the precisely—defined and widely—accepted meaning of statistical significance, *sensu stricto*, that has been utilized for decades. To cite two examples, when a climate scientist declares a significant increase in global average temperature over a time period, or a medical scientist declares a significant decrease in cancer tumor size in response to experimental treatment, it is obvious that these claims should be based on the uniformly—accepted definition of what the term 'significant' means in a statistical sense.

Statisticians are well aware of the issues raised in this paper, and introductory texts and courses in applied statistics cover this topic. Nevertheless, as the selected quotations in the introduction, above, indicate, misinterpretation of overlapping standard errors and/or confidence intervals has been, and continues to be, common in a variety of applied disciplines, confirming Johnson's[45] observation that '. . . men more frequently require to be reminded than informed.' This should provide renewed motivation

i. For instructors in applied statistics to be vigilant to the common misunderstandings that non—statisticians can easily fall into once they leave the classroom; and

ii. For reviewers, editors and readers to be more cognizant of what is not an uncommon practice. In particular, we need to emphasize that, 'In general, a gap between bars [around two treatment means] does not ensure significance, nor does overlap rule it out' (Krzywinski & Altman[46]).

## Acknowledgments

Constructive comments by L.A. Brennan, F.C. Bryant, R.D. DeYoung, E.B. Fish, M.L. Friendly, J.L. Grace, J.P. Leonard, H.G. Mansouri and S. Rideout—Hanzak substantially improved this paper. This is manuscript number 16—101 of the Caesar Kleberg Wildlife Research Institute, Texas A&M University—Kingsville, TX. This work was supported by the Caesar Kleberg Wildlife Research Institute, Texas A&M University—Kingsville, Kingsville, TX, USA.

## Conflict of interest

Author declares no Conflict of interest.

## References

1. Fisher RA. RA Fisher Digital Archive. *Adelaide Research & Scholarship*, Australia:1932.
2. Fisher RA. The statistical method in psychical research. *Proceedings of the Society for Psychical Research*. 1929;39:189–192.
3. Little J. Understanding statistical significance: a conceptual history. *Journal of Technical Writing and Communication*. 2001;314:363–372.
4. Stephens PA, Buskirk SW, Del Rio CM. Inference in ecology and evolution. *Trends in Ecology and Evolution*. 2007;224:192–197.
5. Freeman WH. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. *Freeman and Co*. USA: 2001.
6. Lehmann EL. Fisher, Neyman, and the Creation of Classical Statistics. *Springer*.USA: 2011.
7. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods*. 2000;52:241–301.
8. Robinson DH, Wainer H. On the past and future of null hypothesis significance testing. *Journal of Wildlife Management*. 2002;662:263–271.
9. Murtaugh PA. In defense of P values. *Ecology*. 2014;953:611–617.
10. Deming WE. On errors in surveys. *American Sociological Review*. 1944;94:359–369.
11. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ*. 2012;43:279–282.
12. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013;1422:573–603.
13. Cooley RL. Practical Scheffé-type credibility intervals for variables of a groundwater model. *Water Resources Research*. 1999;35:113–126.
14. Burnham KP, Anderson. Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach. 2nd ed. *Springer*, USA: 2002.
15. Madsen, Petersen SL, Roundy BA, et al. Comparison of postfire soil water repellency amelioration strategies on blue bunch wheatgrass and cheatgrass survival. *Rangeland Ecology and Management*. 2012;652:182–188.
16. Beck TJ, Gawlik DE, Pearlstine EV. Community patterns in treatment wetlands, natural wetlands, and croplands in Florida. *The Wilson Journal of Ornithology*. 2013;1252:329–341.
17. Jones T, Kulseth S, Mechtenberg K, et al. Simultaneous evolution of competitiveness and defense: induced switching in Arabis drummondii. *Plant Ecology*. 2006;1842:245–257.
18. Wisdom MJ, Bate LJ. Snag density varies with intensity of timber harvest and human access. *Forest Ecology and Management*. 2008;255:2085–2093.
19. Dwanisa JP, Mohanty AK, Misra M, et al. Novel soy soil based polyurethane composites: fabrication and dynamic mechanical properties evaluation. *Journal of Materials Science*. 2004;39:1887–1890.
20. Huang GB, Ramesh M, Berg T, et al. Labeled Faces in the Wild: A Database for Studying, Face Recognition in Unconstrained Environments. *Faces in Real-Life Images Workshop in European Conference on Computer Vision ECCV*. Germany: 2008.
21. Smyth FL, Nosek BA. On the gender-science stereotypes held by scientists: explicit accord with gender-ratios, implicit accord with scientific identity. *Front Psychol*. 2015:6–415.

22. Monfils MJ, Brown PW, Hayes DB, et al. Breeding bird use and wetland characteristics of diked and undiked coastal marshes in Michigan. *Journal of Wildlife Management*. 2014;78:79–92.

23. Kilgo JC, Vukovich M, Ray HS, et al. Coyote removal, understory cover, and survival of white-tailed deer neonates. *Journal of Wildlife Management*. 2014;78:1261–1271.

24. Mantha S, Thisted R, Foss J, et al. A proposal to use confidence intervals for visual analog scale data for pain measurement to determine clinical significance. *Anesth Analg*.1993;775:1041–1047.

25. Bekele A, Kellman L, Beltrami H. Soil profile $CO_2$ concentrations in forested and clear cut sites in Nova Scotia, Canada. *Forest Ecology and Management*. 2007;242:587–597.

26. Payton ME, Miller AE, Raun WR. Testing statistical hypotheses using standard error bars and confidence intervals. *Communications in Soil Science and Plant Analysis*. 2000;316:547–551.

27. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance. *J Insect Sci.* 2003;3–34.

28. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences. 4th ed. *Stanford Libraries*, USA: 2013.

29. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin*. 1946;26:110–114.

30. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*. 1954;253:484–498.

31. Goldstein H, Healy MJ. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society.* 1995;1581:175–177.

32. D Afshartous, Preston RA. Confidence intervals for dependent data: equating nonoverlap with statistical significance. *Computational Statistics and Data Analysis*. 2010;5410:2296–2305.

33. Meehl GA, Arblaster JM, Fasullo JT, et al. Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nature Climate Change.* 2011;1:360–364.

34. Carmer S, Swanson M. An evaluation of ten pair wise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association.* 1983;68341:66–74.

35. Cherry S. Statistical tests in publications of The Wildlife Society. *Wildlife Society Bulletin.* 1998;264:947–953.

36. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician.* 2001;553:182–186.

37. Berger JO, Delampady M. Testing precise hypotheses. *Statistical Science.* 1987;23:317–352.

38. Lin M, Lucas HC, Shmueli G. Too big to fail: Large sample sizes and the p-value problem. *Information Systems Research*. 2013; 24:906–917.

39. Friendly M. Data ellipses, HE plots and reduced rank displays for multivariate linear models: SAS software and examples. *Journal of Statistical Software.* 2006;176:1–43.

40. Friendly M. HE plots for multivariate general linear models. *Journal of Computational and Graphical Statistics.* 2007;162:421–444.

41. Fox J, Friendly M, Monette G. Visual hypothesis tests in multivariate linear models: the heplots package for R. Canada, 2007;p. 1–24.

42. Cummings G, Finch S. Inference by eye: Confidence intervals and how to read pictures of data. *Am Psychol.* 2005;602:170–180.

43. Cummings G, Williams J, Fidler F. Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics.* 2004;34:299–311.

44. Wolfe JD, Stouffer PC, Seeholzer GF. Variation in tropical bird survival across longitude and guilds: a case study from the Amazon. *Oikos.* 2014;1238:964–970.

45. Johnson S. The Rambler, Edinburgh, Glasgow, Scotland, 1750;1:347.

46. Krzywinski M, Altman N. Use box plots to illustrate the spread and differences of samples. *Nature.* 2013;10:921–922.